IMPORTANCE OF THE DYNAMIC RANGE OF AN ANALYSIS WINDOW FUNCTION FOR PHASE-ONLY AND MAGNITUDE-ONLY RECONSTRUCTION OF SPEECH

Kamil K. Wójcicki and Kuldip K. Paliwal

Signal Processing Laboratory
Griffith University, Nathan, Brisbane, Australia
{K.Wojcicki, K.Paliwal}@griffith.edu.au

ABSTRACT

The short-time Fourier transform (STFT) of a speech signal has two components: the short-time magnitude spectrum and the short-time phase spectrum. It is traditionally believed that the short-time magnitude spectrum plays the dominant role for speech perception at small window durations (20–40ms). However, recent perceptual studies have shown that the short-time phase spectrum can contribute as much to speech intelligibility as the short-time magnitude spectrum. It was observed that the use of the rectangular (non-tapered) analysis window for the computation of the short-time phase spectrum is more advantageous than the use of the Hamming (tapered) analysis window. This paper investigates the effect that the dynamic range of an analysis window has on the intelligibility of speech for phase-only and magnitude-only stimuli. For this purpose, the Chebyshev analysis window with adjustable equi-ripple side-lobes is employed. Two types of magnitude-only stimuli are investigated: random phase and zero phase. It is shown that the intelligibility of the magnitude-only stimuli constructed with zero phase is independent of the dynamic range of the analysis window, while the random phase stimuli are intelligible only for analysis windows with high dynamic range. This study also shows that for low dynamic range analysis windows, the short-time phase spectrum at small window durations (20–40ms) contributes as much as to speech intelligibility as the short-time magnitude spectrum.

Index Terms— Short-time magnitude spectrum, short-time phase spectrum, speech processing

1. INTRODUCTION

Although speech is non-stationary, it can be assumed quasi-stationary and, therefore, can be processed through a short-time Fourier analysis. Note that the modifier 'short-time' implies a finite-time window over which the properties of speech may be assumed stationary; it does not refer to the actual duration of the window.[1] The short-time Fourier transform (STFT) of a speech signal $s(t)$ is given by

$$S(f,t) = \int_{-\infty}^{\infty} s(\tau) w_a(t-\tau) \mathrm{e}^{-j2\pi f\tau} d\tau, \qquad (1)$$

where $w_a(t)$ is an analysis window function of duration T_w. In speech processing, the Hamming window function is typically used and its width is normally 20–40ms. The short-time Fourier spectrum, $S(f,t)$, can be expressed as

$$S(f,t) = |S(f,t)| \mathrm{e}^{j\psi(f,t)}, \qquad (2)$$

where $|S(f,t)|$ is the short-time magnitude spectrum and $\psi(f,t) = \angle S(f,t)$ is the short-time phase spectrum. The signal $s(t)$ is completely characterised by its magnitude and phase spectra.

In our previous paper [1],[2] we investigated the intelligibility resulting from magnitude-only and phase-only stimuli. The rectangular and Hamming analysis windows were used. For small window durations (20–40ms), the rectangular window produced phase-only stimuli with comparable intelligibility to that of magnitude-only stimuli;[3] while, for the Hamming window, the intelligibility of magnitude-only stimuli was found to be significantly higher than that of phase-only stimuli. These findings pose an interesting question as to what property makes the rectangular window better suited for the estimation of the short-time phase information. Is it the difference in attenuation of the highest side-lobe with respect to (w.r.t.) the main-lobe, or is it some other property? Therefore, our main aim in this paper is to find the reason why the phase-only stimuli provides us with better intelligibility when the rectangular window is used instead of the Hamming window. For this purpose, we have selected the Chebyshev[4] window in which we can systematically change the dynamic range[5]. In our experiments we use a small window duration (32ms), which is commonly employed in both speech recognition and speech processing applications.

In the previous study [1], the following two methods were used to eliminate the phase information in magnitude-only stimuli construction. In the first method, the phase information was removed by setting the short-time phase spectrum values to zero, while in the second method, the phase information was removed by randomising the phase spectrum values. We found that the random phase gave us slightly better results, hence we reported our results for magnitude-only random phase stimuli. Over the course of the present study, however, we found that when the dynamic range of the Chebyshev analysis window was very small (5–10dB), the magnitude-only stimuli constructed with zero phase gave much better results than the ones constructed with random phase. This motivated us to investigate the effect that the dynamic range of an analysis window has on the intelligibility of the magnitude-only stimuli constructed with random phase and zero phase.

[1] We use the qualitative terms 'small' and 'large' to refer to the duration.

[2] When we refer to 'our previous paper', we mean the study reported by our group from Signal Processing Laboratory at Griffith University.

[3] The rectangular window has also been recommended in the literature for computation of group delay spectrum, which is a frequency derivative of phase spectrum [2].

[4] Chebyshev window is also know in the literature as Dolph-Chebyshev window.

[5] By the 'dynamic range' we mean attenuation of the side-lobe level w.r.t. to the main-lobe.

Consequently, the aim of the present study is two fold. First, we investigate the effect of the dynamic range of an analysis window on the intelligibility of magnitude-only stimuli constructed with random phase and zero phase. Second, we investigate the effect of the dynamic range of an analysis window on the intelligibility of phase-only stimuli and compare it with magnitude-only stimuli constructed with zero phase since this approach produces consistently better results than the random phase approach.

This paper is organised as follows. In Section 2, we detail the analysis-modification-synthesis (AMS) procedure used to generate the stimuli files. Description of the perception experiments is given in Section 3. The results and discussion are presented in Section 4.

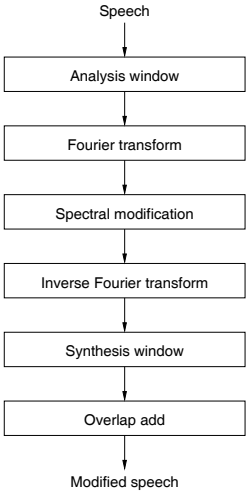

Figure 1: *Speech analysis-modification-synthesis (AMS) procedure used for generation of stimuli files.*

2. ANALYSIS-MODIFICATION-SYNTHESIS PROCEDURE

The main aim of this study is to determine the relative contribution that the magnitude and phase spectra provide towards speech intelligibility.[6] Accordingly, stimuli retaining only the magnitude or phase information are constructed. For this purpose, an analysis-modification-synthesis (AMS) procedure, shown in Fig. 1, is used. In the AMS framework the speech signal is divided into overlapped frames of small duration. The frames are then windowed using an analysis window, $w_a(t)$, followed by Fourier analysis, and spectral modification. The spectral modification stage is where only the magnitude or phase information is retained. For example, to construct phase-only (PO) stimuli the magnitude spectrum is set to unity while the phase spectrum is left unchanged, resulting in

$$\hat{S}(f,t) = \mathrm{e}^{j\psi(f,t)}, \tag{3}$$

where $\hat{S}(f,t)$ is the modified STFT. The stimuli, $\hat{s}(t)$, is then constructed by taking the inverse STFT of the $\hat{S}(f,t)$, followed by an overlap-add (OLA) synthesis [3, 4, 5, 6].[7] The resulting signal contains all of the information about the short-time phase spectra contained in the original signal $s(t)$, but has no information about the short-time magnitude spectra. Similarly, for generation of magnitude-only stimuli we retain only the magnitude information of each frame by removing the phase spectrum information. There are two approaches for removal of phase spectrum information from $S(f,t)$. In the first approach, the phase spectrum values for each frame are set to zero, and so the modified STFT is

$$\hat{S}(f,t) = |S(f,t)|. \tag{4}$$

We refer to the resulting stimuli as magnitude-only zero phase (MOZP) stimuli. In the second approach, the phase spectrum values are randomised for each frame. The resulting modified STFT is given by

$$\hat{S}(f,t) = |S(f,t)|\mathrm{e}^{j\phi}, \tag{5}$$

where ϕ is a random variable uniformly distributed between 0 and 2π.[8] We refer to the resulting stimuli as magnitude-only random phase (MORP) stimuli.

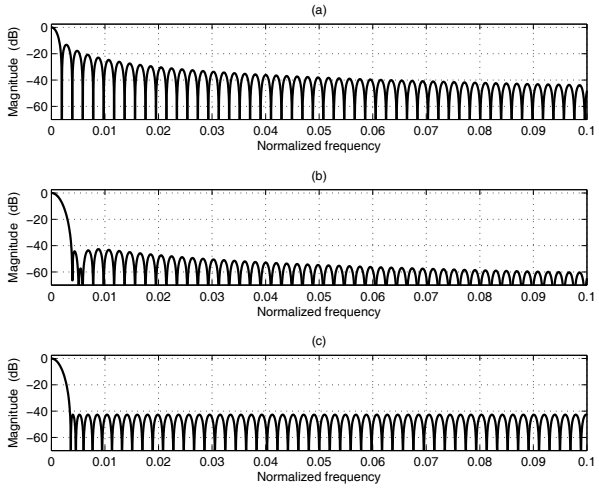

Figure 2: *Magnitude spectrum characteristics of three window functions: (a) rectangular window; (b) Hamming window; (c) Chebyshev window (with chosen side-lobe attenuation of 42.7dB). Here, the window length used is N=512 samples, and the FFT length is nfft=64N.*

Recent perceptual studies [7, 8, 9, 1, 10] employed Hamming and rectangular analysis windows. They showed that the rectangular window is beneficial for estimation of short-time phase spectrum over small window durations (20–40ms). In this study, our goal is to investigate the effect of the analysis window in more detail. In particular the effect that the dynamic range of the analysis window has on speech intelligibility of magnitude-only and phase-only stimuli is investigated. For this purpose, we employ the Chebyshev window [11] characterised by adjustable equi-ripple side-lobe attenuation. The comparison of spectral responses of the aforementioned window functions is shown in Fig. 2. Note that the side-lobes of the Hamming and rectangular windows 'roll-off' with frequency at different rates, while they remain constant for the Chebyshev window. As a consequence, it can be difficult to

[6] Throughout our discussions, when referring to phase or magnitude spectrum, the use of short-time Fourier transform (STFT) over small window durations (20–40ms) is implied, unless otherwise stated.

[7] In the following experiments we use Griffin and Lim's reconstruction method [6].

[8] Note that when constructing the random phase spectrum, the antisymmetry property of phase spectrum should be preserved.

draw meaningful comparisons between them. For this reason, in this paper, we reserve the term 'dynamic range' exclusively for the Chebyshev analysis window; i.e. we use it to refer to the attenuation of the main-lobe w.r.t. to the (uniform) side-lobe level and not to the more typically used, broader, category of main-lobe w.r.t. the highest side-lobe. For the synthesis window we employ the modified Hanning window [6].

3. HUMAN SPEECH PERCEPTION EXPERIMENTS

3.1. Recordings

Six stop consonants, /b, d, g, k, p, t/, were selected for the recognition task. Each consonant was placed in a vowel-consonant-vowel (VCV) context within the 'Hear aCa now' carrier sentence.[9] The recordings were carried out in a silent room using a SONY ECM-MS907 microphone. Four speakers were used, two males and two females. Six recordings per speaker were made, giving a total of 24 recordings. Each recording lasted approximately three seconds, including leading and trailing silence portions. All recordings were sampled at 16 kHz with 16-bit precision.

3.2. Stimuli

The recordings were processed using the analysis-modification-synthesis procedure detailed in Section 2. This was performed to retain only phase or magnitude information. Further, to determine the effect that the dynamic range of an analysis window has on speech intelligibility, nine different window functions were employed. Seven Chebyshev windows with the dynamic ranges from 5 to 65dB, with 10dB increments were used. To enable comparison with previous studies the rectangular and Hamming windows were also included. Small window durations, T_w=32ms, were used throughout. The frame shift was set to $T_w/8$ to minimise aliasing. The FFT analysis length was set to $2N$, where N is the number of samples in each frame. The resulting stimuli can be grouped as follows: 1) phase-only (PO), 2) magnitude-only zero phase (MOZP), and 3) magnitude-only random phase (MORP). The original recordings (reconstructed without modification) were also included. Overall, 28 different treatments were applied to the 24 recordings, resulting in the total of 672 stimuli files.

3.3. Subjects

For listeners, we used 12 English speaking volunteers, with normal hearing. None of the listeners participated in the recording of the stimuli.

3.4. Procedure

The perception tests were conducted in isolation, over a single session, in a quiet room. The task was to identify each carrier utterance as one of the six stop consonants. The listeners were presented with seven labelled options on a digital computer, with the first six corresponding to the six stop consonants and the seventh being the null response. The subjects were instructed to choose the null response only if they had 'no idea' as to what the embedded consonant might have been. The stimuli audio files were played in a randomised order and presented over closed circumaural headphones (SONY MDR-V500) at a comfortable listening level. Prior to the actual test, the listeners were familiarised with the task in a short practice session. The entire sitting lasted approximately 90 minutes with numerous five minute breaks. The responses were collected via a keyboard. No feedback was given.

[9]For example, for the consonant /b/, the utterance is 'Hear aba now'.

Table 1: *Average consonant intelligibility scores (%) with corresponding standard deviations for PO, MOZP, and MORP stimuli. Results for a range of analysis window types are shown.*

	Treatment Groups		
Analysis Window	PO	MOZP	MORP
Original	99.7 ± 1.2	99.7 ± 1.2	99.7 ± 1.2
Chebyshev 5dB	98.6 ± 2.1	97.2 ± 4.8	0.0 ± 0.0
Chebyshev 15dB	99.3 ± 1.6	98.6 ± 2.1	3.8 ± 6.3
Chebyshev 25dB	97.9 ± 2.8	99.7 ± 1.2	50.7 ± 11.2
Chebyshev 35dB	95.1 ± 3.9	99.0 ± 1.9	82.6 ± 9.4
Chebyshev 45dB	93.8 ± 3.8	99.0 ± 1.9	94.1 ± 2.8
Chebyshev 55dB	89.9 ± 4.5	98.3 ± 2.1	98.6 ± 2.1
Chebyshev 65dB	81.6 ± 10.6	97.6 ± 2.8	99.7 ± 1.2
Rectangular	96.5 ± 2.4	98.6 ± 2.7	95.8 ± 4.4
Hamming	67.7 ± 6.2	96.9 ± 2.6	98.6 ± 2.1

4. RESULTS AND DISCUSSION

The results of our experiments are shown in Table 1. The results in the first row refer to the intelligibility of the original stimuli. The next seven rows show the intelligibility results for the Chebyshev window at different dynamic ranges. Rows nine and ten give the results for the rectangular and Hamming windows, respectively. We separate our discussion of the results into two sections. First, we address the issue of magnitude-only stimuli construction by evaluating the intelligibility of magnitude-only stimuli constructed with two different methods: random phase and zero phase. Second, we compare the phase-only stimuli with magnitude-only zero phase stimuli, since the zero phase stimuli produces consistently better results than the random phase stimuli.

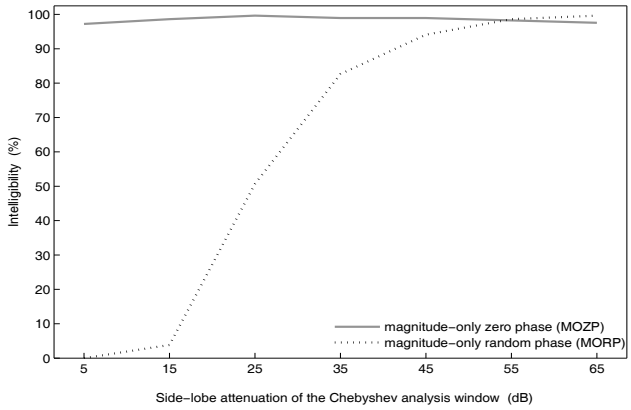

Figure 3: *Consonant identification performance (or, intelligibility) (%) as a function of the dynamic range of the Chebyshev analysis window for MOZP stimuli (solid line) and MORP stimuli (dotted line).*

4.1. Intelligibility of MOZP stimuli versus MORP stimuli

The comparison of intelligibility of MOZP and MORP stimuli as a function of the dynamic range of the Chebyshev analysis window is shown in Fig. 3. Based on this comparison, as well as on the results from Table 1, the following observations can be made:

1. The intelligibility of the MOZP stimuli is consistently high and does not depend on the dynamic range of the analysis window. Within-subjects repeated measures ANOVA test confirms this observation ($F[8, 88]$=1.865, p=0.076). On the other hand, the dynamic range of an analysis window has a significant effect on the intelligibility of the MORP stimuli ($F[8, 88]$=712.04, $p<0.01$).
2. For low dynamic range analysis windows (i.e., Chebyshev $\leq$45dB) the intelligibility of the MOZP stimuli is significantly higher than the intelligibility of the MORP stimuli, while, for high dynamic range (i.e. Chebyshev $>$45dB) the difference in the intelligibility of MOZP and MORP stimuli is insignificant. This was confirmed by paired t-tests at a $p<0.01$ level of significance.

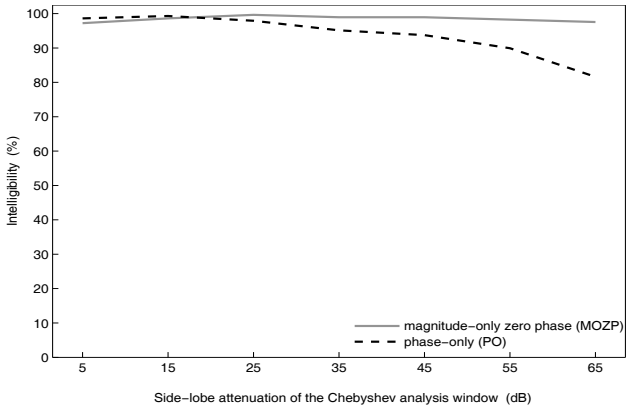

Figure 4: *Consonant identification performance (or, intelligibility) (%) as a function of the dynamic range of the Chebyshev analysis window for MOZP stimuli (solid line) and PO stimuli (broken line).*

4.2. Intelligibility of MOZP stimuli versus PO stimuli

The intelligibility of MOZP and PO stimuli as a function of the dynamic range of the Chebyshev analysis window is compared in Fig. 4. The following observations can be made from Table 1 and Fig. 4:

1. The dynamic range of the analysis window has a insignificant effect on the intelligibility of MOZP stimuli ($F[8, 88]$= 1.865, p=0.076); while it has a significant effect on the intelligibility of the PO stimuli ($F[8, 88]$=68.365, $p<0.01$).
2. For low dynamic range analysis windows (Chebyshev $\leq$35dB) the intelligibility of MOZP stimuli does not significantly differ from the intelligibility of PO stimuli. However, MOZP stimuli has a significantly higher intelligibility than PO stimuli for analysis windows with large dynamic range (Chebyshev $>$35dB). This was confirmed by paired t-tests at a $p<0.01$ level of significance.

These observations can be explained as follows. The phase spectrum can be computed using arctangent function (four quadrant version) as

$$\psi(f,t) = \arctan\left(\frac{\mathrm{Im}\{S(f,t)\}}{\mathrm{Re}\{S(f,t)\}}\right), \quad (6)$$

where $\mathrm{Re}\{\cdot\}$ and $\mathrm{Im}\{\cdot\}$ denote real and imaginary parts, respectively. From (6), it is evident that when $|S(f,t)|$ is very small the argument of arctan function will be of $(\frac{0}{0})$ form; hence the phase estimates will not be numerically reliable. Such small values occur when the dynamic range of $w_a(t)$ (and thus $|S(f,t)|$) is large. On the other hand, low dynamic range analysis windows produce better behaved phase spectra estimates, since higher side-lobes imply that $|S(f,t)|$ is always relatively large.

5. CONCLUSION

In this paper, the importance of the dynamic range of an analysis window on speech intelligibility of short-time magnitude and phase spectra was investigated. The stimuli were constructed by retaining only the short-time magnitude or phase spectrum information, at short window durations (32ms). It was shown that the intelligibility of magnitude-only stimuli constructed with zero phase is independent of the dynamic range of the analysis window, while random phase stimuli are intelligible only for analysis windows with high dynamic range. This study also shows that for low dynamic range analysis windows the phase spectrum contributes as much to speech intelligibility as the magnitude spectrum.

6. REFERENCES

[1] K.K. Paliwal and L.D. Alsteris, "On the usefulness of STFT phase spectrum in human listening tests," *Speech Communication*, vol. 45, no. 2, pp. 153–170, Feb. 2005.

[2] N.S. Reddy and M.N.S. Swamy, "Derivative of phase spectrum of truncated autoregressive signals," *IEEE Trans. Circ. Systems*, vol. CAS-32, no. 6, pp. 616–618, June 1985.

[3] J.B. Allen and L.R. Rabiner, "A unified approach to short-time Fourier analysis and synthesis," *Proc. IEEE*, vol. 65, no. 11, pp. 1558–1564, Nov. 1977.

[4] R.E. Crochiere, "A weighted overlap-add method of short-time Fourier analysis / synthesis," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, no. 2, pp. 99–102, Feb. 1980.

[5] M.R. Portnoff, "Short-time Fourier analysis of sampled speech," *IEEE Trans. Acoust., Speech and Signal Processing*, vol. ASSP-29, no. 3, pp. 364–373, 1981.

[6] D.W. Griffin and J.S. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-32, no. 2, pp. 236–243, Apr. 1984.

[7] K.K. Paliwal and L.D. Alsteris, "Usefulness of phase spectrum in human speech perception," in *Proc. Eurospeech*, Geneva, Switzerland, Sept. 2003, pp. 2117–2120.

[8] K.K. Paliwal, "Usefulness of phase in speech processing," *Proc. IPSJ Spoken Language Processing Workshop, Gifu, Japan*, pp. 1–6, Feb. 2003.

[9] L.D. Alsteris, *Short-time phase spectrum in human and automatic speech recognition*, Ph.D. thesis, Griffith University, Brisbane, Australia, Aug. 2005.

[10] L.D. Alsteris and K.K. Paliwal, "Further intelligibility results from human listening tests using the short-time phase spectrum," *Speech Communication*, vol. 48, no. 6, pp. 727–736, June 2006.

[11] F.J. Harris, "On the use of windows for harmonic analysis with the discrete Fourier transform," *Proc. IEEE*, vol. 66, no. 1, pp. 51–83, Jan. 1978.